A Multi-Lingual Architecture for Building a Normalised Conceptual Representation from Medical Language*

P. Zweigenbaum, Ph.D., B. Bachimont, Ph.D., J. Bouaud, Ph.D., J. Charlet, Ph.D., J.F. Boisvieux, M.D., Ph.D.

DIAM — INSERM U.194 and
SIM, Service d'Informatique Médicale, Assistance Publique – Hôpitaux de Paris
{pz ,bb, jb, jc, jfb}@biomath.jussieu.fr

The overall goal of MENELAS *is to provide better access to the information contained in natural language patient discharge summaries (*PDSs*), through the design and implementation of a prototype able to analyse medical texts. The approach taken by* MENELAS *is based on the following key principles: (i) to maximise the usefulness of natural language analysis and the usability of its results, the output of natural language analysis must be a normalised conceptual representation of medical information; and (ii) to maximise the reuse of resources, language analysis should be domain-independent and conceptual representation should be language-independent. This paper discusses the results obtained and the issues raised when implementing these principles during the project.*

INTRODUCTION

Medical language processing is now a well-developed field of research (see, *e.g.*, [1, 2]), and a number of prototypes and systems have been built for various languages and purposes (*e.g.*, [3, 4, 5, 6]). In these systems, addressing a new language generally requires the development of the corresponding linguistic (lexical, morphological, syntactic) knowledge. Starting from the observation that such general linguistic resources are available for more and more languages, we have studied an architecture which can ease their reuse for medical language processing.

The output of a natural language processor can be close to the input language [1], aim at a more conceptual representation [4, 5, 6], or can also use a standardised vocabulary (*e.g.*, [3]). We defend the production of a normalised conceptual representation, a sort of "interlingua" with formal generative capabilities. The advantage of such a representation is to be abstracted away from the initial linguistic form, which facilitates subsequent computation and thus the use of the representation for a variety of purposes. For instance, coding into a standardised vocabulary then becomes a more formal transcoding task than direct coding from natural language. Aiming for such a normalised conceptual representation is by no means a trivial task, and raises the issue of its production from natural language input. The very design of the representation is itself a fundamental point, which we have addressed elsewhere [7, 8].

In this paper, we first recall the general objectives of the MENELAS project and its architectural principles. We then describe the implemented prototype and its results. We finally discuss the issues raised and the experience gained from the project.

PROJECT OBJECTIVES

The overall objectives of MENELAS [9] fall into two categories: target, user-level services which will be created or enhanced by extracting medical information from free text, and advances in the enabling medical language processing technology. The larger issues and effort needed lie in the development of the enabling technology. Nevertheless, the project has taken care to include user-level service demonstration modules in its design. This helped to show the end-user benefits of the work and to focus the R&D work on the strategic points to be addressed to achieve such user-level services. Target, user-level services are organised around two main axes; a further study of clinicians' requirements for NLP-related tools, performed since then in the EU DOME project (MLAP 63221), confirmed and elaborated on these services:

Coding: the automatic assignment of codes to each PDS. In most European countries, each patient stay must be indexed with codes from national or international nomenclatures and classifications. This coding task is imposed on health care professionals for external reasons, and constitutes extra work with little direct benefit. In this context, such a service can save much time and effort. In MENELAS, the *production of ICD-9-CM codes from* PDS *texts* has been addressed.

Information retrieval: Patient cases can be retrieved on the basis of the full contents of the discharge summary. This may be used for clinical and research purposes: (*i*) the clinical staff may access the stored PDSs to *retrieve a set of patients having specific characteristics*, which allows some form of human case-based reasoning; (*ii*) the same facility allows a limited *test of research hypotheses*,

*This work has been partly supported by the European Union project MENELAS (AIM 2023).

on the basis of the available patient sample. The analysed contents of the texts may also be used to feed activity management tools.

These services can be implemented as modular subsystems, and be realised gradually. The project has focussed on the production of ICD-9-CM codes and a query-based retrieval functionality. The test medical domain is coronary diseases.

At the technical level, the goals of the project were to adapt and develop advanced text analysis techniques to build a canonical representation of the medical information contained in PDSs. We now turn to the key points of the chosen approach.

ARCHITECTURAL PRINCIPLES

Normalised core representation

The relevant medical information contained in texts is extracted and structured according to a normalised, conceptual representation. This representation is abstracted away from its original linguistic form: it is robust with respect to the variability and flexibility of language and across different languages. It provides a normalised, systematic informational basis to support smoothly a palette of user-level services: classification or nomenclature codes (here, ICD-9-CM) can be extracted from it; queries can be matched against it to retrieve specific information contained in a PDS. A direct benefit is that the components which implement these services do not need to have any knowledge of the initial natural language used in the input text, nor do they have to cope with natural language problems such as synonymy or paraphrase. We do not claim however that such a representation can be universal. It necessarily depends on a point of view, imposed by the task(s) which will use the information. This approach contrasts with that of [1], whose target representation contains words of the source language, so that a subsequent coding module still needs to rely on linguistic knowledge [3].

Common, multiple-language architecture

The architecture should maximise language-independent, shared components and knowledge bases. The twelve partners of MENELAS represent three linguistic groups: French, English and Dutch, addressed during the project. The core representation, and as a result, the code generation component and the retrieval component, are language-independent, as well as the medical knowledge description they share. While sharing with the Geneva approach [4] similar goals of maximal reusability across languages, the architecture adopted by MENELAS is open to inclusion of already existing language-specific components to facilitate extension to other natural languages.

Reusing existing linguistic resources

Despite the growing availability of general linguistic resources (lexica, parsers, grammars, etc.), most natural language processing systems are gradually built by one team with their own background material. In contrast, MENELAS builds on previous work performed by the different partners either on medical language processing, both on French [6] and on Dutch [10], or on the analysis of general French and English [11, 12]. The latter, in particular, were large-coverage parsers with lexica over 40,000 words. Complementary developments were performed where necessary.

Full text processing

PDS texts may contain both full prose paragraphs, with grammatically complex sentences, and shorter noun phrase statements. Non-trivial syntactic phenomena, such as embedded controlled clauses or negation, can have a substantial impact on text meaning. It is hence difficult to dispense from taking them into account. Parsers with advanced syntactic knowledge, which attempt full sentence parsing, are therefore useful here — and are more and more available in various languages. This approach can be compared with that of [1, 5], who use 'conventional' parsing technology in their systems; it contrasts with that of [4], who use an original non-grammar-based analysis technique.

In-depth, knowledge-based approach

MENELAS adopts a knowledge-based approach to natural language understanding. This pushes forward previous work such as [6] or [5], trying to make the most of both syntactic and conceptual knowledge. Domain knowledge is all the more necessary as a higher level of canonisation from free text is desired. It depends on the application domain, but is not specific to a language. It is expressed in the Conceptual Graph (CG) formalism [13], which is gradually emerging as a knowledge representation standard in medical systems [4, 5].

IMPLEMENTATION AND RESULTS

According to the above objectives, MENELAS has a two-part organisation (Figure 1). The heart of MENELAS is the document **analysis system**, which analyses a PDS and stores it in a database as a set of normalised, conceptual representations. These representations hold the informational basis for the target services. While a large range of services could exploit the information extracted by the analysis system, the project has concentrated on two **target services**: (*i*) *the coding service* produces nomenclature codes; (*ii*) *the consultation service* allows physicians and management staff to retrieve PDSs that satisfy content-based criteria.

It is a non-trivial task to adapt large, complicated existing systems, processing different natural languages and implemented in different programming languages, to fit into a new, shared architecture. Careful design of a modular architecture and of

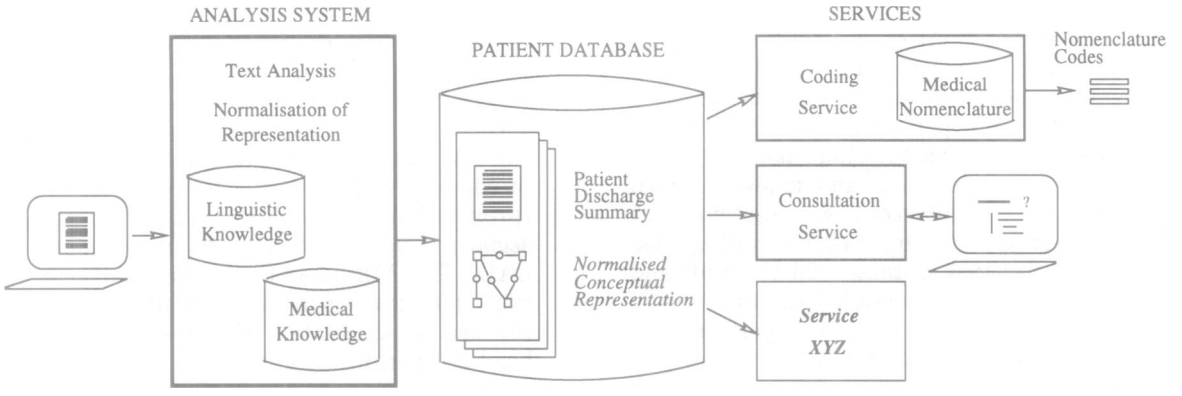

Figure 1: The MENELAS architecture, centered on a shared, normalised representation of medical information.

its data interchange formats was therefore one of the first tasks of the project (see Figure 2). An existing language processing chain up to *morpho-syntactic analysis* can be integrated provided its output follows the Annotated Parse Tree (APT) format specified during the project; this was done for French and Dutch [14]. A processing chain up to *semantic analysis* can be connected if its output is in the CG exchange format specified during the project; this was the case for English [15].

Given space constraints, we can only list the main software components, tools, and knowledge bases output by the project. *Language analysis components* for French, English [15] and Dutch [14]; language-generic pragmatic analyser and conceptual graph toolbox [16]; *Code generator component* parameterised for ICD-9-CM [17]; *Information retrieval components* [18]; *Medical knowledge bases:* language-independent ontology (1800 elementary concepts, over 250 relations) and models (over 500 reference models) for the domain of coronary diseases [7, 19]. These components have been implemented in PROLOG, LISP, C++ and C, and run on Unix; the consultation user interface runs on PC.

Given input coming from different languages, or even within a single language, starting from variable surface forms, the analysis system must provide a canonical representation of the informational content of input utterances. We have therefore designed and implemented a method which ensures that all representations that are built conform to the norm. The method consists in projecting linguistic relations to reference models of the domain, heuristically selecting the conceptual path which best links the corresponding notions, with the help of conceptual preferences associated to prepositions and other grammatical relations. The procedure can assign identical projections to linguistic paraphrases. At the same time, it can discriminate linguistically similar expressions which have different meanings.

The service modules can then take advantage of

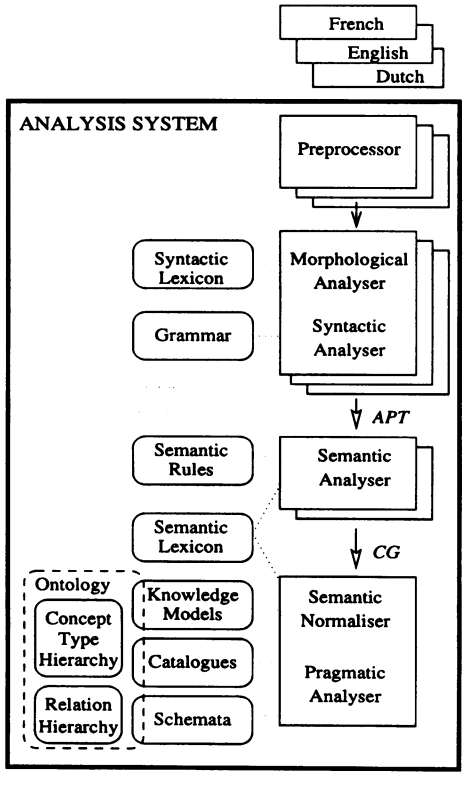

Figure 2: Document Analysis System architecture.

this canonised form of information. The Coding Service, *e.g.*, has no knowledge of natural language and no need to know the specific ICD-9-CM code labels for French, English and Dutch. Its criteria for assigning a given code are directly expressed in terms of the normalised conceptual representation.

Preliminary evaluation [9] shows that the existing prototype displays promising results for automatically encoding PDSs into ICD-9-CM and for full-text information retrieval from natural language PDSs. Partial tests (up to syntactico-semantic analysis) have been performed on a large PDS sample (475 English texts and 100 French texts). In-depth tests of the whole analysis chain were per-

formed on a smaller, 37 text sample, representing 393 sentences of up to 96 words [20]. they consisted in comparing the user-level service results (coding and query answering) obtained by the system with the performance of health care professionals performing the same tasks: given a PDS text, assign codes to it and answer a fixed questionnaire. Overall *recall* was measured at 48 %, overall precision at 63 % on the coding task; the questionnaire task obtained 66 % recall and 77 % precision. A more detailed examination of the results shows that MENELAS can perform better than traditional information retrieval methods on a variety of queries, such as knowledge-based queries and complex queries, and can help human coders to improve coding consistency [20].

The semantic normalisation procedure can successfully discriminate between similar surface forms (for space reasons, we only provide a gloss for CG representations): *e.g.*, *'admitted for angina pectoris'* (reason) *vs 'admitted for angioplasty'* (goal); *'angioplastie de l'IVA'* (angioplasty acting on an artery segment part of the LAD) *vs 'angioplastie de la lésion'* (on an artery segment in a stenosed state) *vs 'angioplastie de Monsieur X'* (on an artery segment part of a human being — male, name is *'X'*). It can also assign identical representations to variant forms of the same expressions: *e.g.*, *'angioplastie sur la sténose circonflexe'* = *'angioplastie sur la circonflexe'* = *'angioplastie circonflexe'*; across languages, *'Patient [...] hospitalisé pour angor [...]'* = *'This [...] patient was admitted for [...] angina pectoris'* = *'De [...] patient werd gehospitaliseerd voor [...] angor'*.

DISCUSSION

The experience of the project sheds some light on the principles set at the beginning of the work. Whereas an ideal goal could be to build a normalised core representation that would be as general as possible, we confirmed that the specification of such a representation necessarily depends to a fair extent on the point of view imposed by the client services and the *target information* they require. Any generality of the obtained canonical form beyond these tasks can only be putative. This target information was specified with the help of users; it was then taken as a reference by knowledge engineers during the design of the normalised representation and system knowledge bases.

By construction, the semantic normalisation procedure always maps to representations which conform to the reference models. *Discrimination* depends on the availability of alternate paths in the models, hence in their richness and accuracy. It also depends on the correct assignment of conceptual preferences to grammatical relations. In experiments performed so far, we were able to obtain good discrimination results. *Identification* depends not only on the selection of a fitting conceptual description for a given input, but also on the recovery around this conceptual description of complementary information through inference. This is necessary, *e.g.*, to be able to identify *'sténose de l'IVA proximale'* (stenosis of the proximal LAD) with *'sténose proximale de l'IVA'* (proximal stenosis of the LAD) and *'sténose du segment I de l'IVA'* (stenosis of segment I of the LAD). The MENELAS pragmatic analyser includes such inference mechanisms in its design through the use of schemata. However, too little work has been performed until now on schema development, and identification performance could therefore greatly be improved.

The bulk of the work on MENELAS knowledge bases concerned (*i*) at a fundamental, conceptual level, the ontology: more than 2,000 atomic types and 6,000 of their instances; (*ii*) at a linguistic level, mappings from lemmas (words) to conceptual representations, for more than 1,000 lexical items. Such work would have been impossible without principled development guidelines. Successive design revisions along the project have been motivated by the progressive elaboration of such guidelines [7, 8]. They have resulted in greater consistency and effectiveness of the MENELAS knowledge bases. There is no doubt complementary principles are still needed; further experimentation will lead to continued progress.

The MENELAS architecture successfully hosted existing French, English and Dutch language processing components, allowing each analysis chain to share the domain-specific knowledge and language-independent components and services. One should not underestimate the cost of adapting and tailoring reused existing resources. Nevertheless, project partners reinjected parsers with large coverage lexica and grammars with a much better benefit/cost ratio than would have resulted from redescribing linguistic knowledge anew for each language addressed. Thanks to their large coverage, the full-text processing parsers used were able to deal at once with a large proportion of the PDS sentences, displaying a wide range of syntactic phenomena. In complement, they also included usual techniques (chart parsing) for dealing with incomplete parses, due, *e.g.*, to sentences with a non-standard syntax. This robustness could be further improved by taking advantage of the domain knowledge available to the language-independent components of the system. However, it is not yet clear how such extensions could be proposed to existing parsers in a generic way.

Finally, the in-depth, knowledge-based approach is really linked to the task: representing correctly and univocally the target information, starting from the variability of natural language. The com-

plexity of this approach cannot be lowered without changing this task.

Acknowledgements
We wish to thank all the hospital departments which provided us corpora, and our colleagues and partners in the MENELAS project. However, this paper expresses views which should not be taken as representing those of the whole consortium.

Main project partners: Pierre Dujols (Groupe Linguistique, Informatique et Médecine, Montpellier, F), Marius Fieschi, Françoise Volot, Nigel Strang (CERTIM, Service de l'Information Médicale, Marseille, F), Pierre Le Beux, Denis Delamarre, Anita Burgun (Laboratoire d'Informatique Médicale, Rennes, F), Mohamed Ben-Saïd (INSERM U.194, Paris, F), Mark Keane, Brenda Nangle, Stephen Flinter (Trinity College Dublin, Dublin, IRL), Brendan McAllister, Stephen Mc Namara, Kevin O'Sullivan, Roy Johnston, Margaret McDermott (Irish Medical Systems, Dublin, IRL), Bridie Kelly (Royal Victoria Hospitals, Belfast, UK), Jos L. Willems†, Peter Spyns (Katholieke Universiteit Leuven, Leuven, B), Antoine Ogonowski (GSI-Erli, Charenton, F), Thierry Guillotin, Jean Fargues, Marie-Claude Landau, Aurel Bradea (IBM, Paris Scientific Center, Paris, F), Marc Moens, Greg Whittemore, Claire Grover, Andrei Mikheev (LTG, HCRC Language Technology Group, Edinburgh, UK).

References

1. Sager N, Friedman C, and Lyman MS, eds. *Medical Information Processing — Computer Management of Narrative Data.* Addison Wesley, Reading, Mass., 1987.
2. Scherrer JR, Côté RA, and Mandil SH, eds. *Computerised Natural Medical Language Processing for Knowledge Engineering*, Amsterdam, 1989. North-Holland.
3. Sager N, Lyman M, Nhàn NT, and Tick LJ. Medical language processing: Applications to patient data representation and automatic encoding. Meth Inform Med 1995;34(1).
4. Baud RH, Rassinoux AM, Wagner JC, et al. Representing clinical narratives by Sowa's Conceptual Graphs. Meth Inform Med 1995;34(1).
5. Schröder M. Knowledge-based processing of medical language: A language engineering approach. In: Proceedings of GWAI'92, Bonn, D. September 1992:190–9.
6. Cavazza M and Zweigenbaum P. A semantic analyzer for natural language understanding in an expert domain. App Artif Intell 1994;8(3):425–53.
7. Zweigenbaum P, Bachimont B, Bouaud J, Charlet J, and Boisvieux JF. Issues in the structuring and acquisition of an ontology for medical language understanding. Meth Inform Med 1995;34(1).
8. Bouaud J, Bachimont B, Charlet J, and Zweigenbaum P. Methodological principles for structuring an "ontology". In: IJCAI'95 Workshop on "Basic Ontological Issues in Knowledge Sharing", August 1995.
9. Zweigenbaum P. MENELAS: an access system for medical records using natural language. Comput Meth Prog Biomed 1994;45:117–20.
10. Spyns P and Adriaens G. Applying and improving the restriction grammar approach for Dutch patient discharge summaries. In: Zampolli A, ed, COLING'92, 1992:1264–8.
11. Fargues J, Landau MC, Dugourd A, and Catach L. Conceptual graphs for semantics and knowledge processing. IBM Journal of Research and Development 1986;30(1):70–9.
12. Grover C, Carroll J, and Briscoe T. *The Alvey Natural Language Tools Grammar.* University of Cambridge, Computer Laboratory, Cambridge, 1992. 4th Release.
13. Sowa JF. *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley, London, 1984.
14. Spyns P and Willems JL. Dutch medical language processing: Discussion of a prototype. In: Greenes RA, Peterson HE, and Protti DJ, eds, Proc MEDINFO 95, Vancouver. 1995:37–40.
15. Whittemore G. The MENELAS English natural language understander: Natural language understanding in the medical domain. In: Proceedings of The First World Congress on Computational Medicine, Public Health, and Biotechnology, Austin, TX. 1994.
16. Bouaud J. Un système de production à base de graphes conceptuels; application dans un système de compréhension de textes. In: PRC-GDR IA , ed, Journées "Graphes Conceptuels", mars 1994.
17. Delamarre D, Burgun A, Seka LP, and Le Beux P. Automated coding system of patient discharge summaries using conceptual graphs. Meth Inform Med 1995. In press.
18. Nangle B and Keane MT. Effective retrieval in hospital information systems. Artif Intell Med 1994;6(3):207–28.
19. Volot F, Zweigenbaum P, Bachimont B, et al. Structuration and acquisition of medical knowledge: Using UMLS in the Conceptual Graph formalism. In: Proc 17th Annu Symp Computer Applications in Medical Care, Washington. Mc Graw Hill, November 1993:710–4.
20. Zweigenbaum P, Bachimont B, Bouaud J, et al. MENELAS final report. Deliverable report AIM-MENELAS 17, DIAM-SIM/INSERM U.194, 1995.